1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS
STANDARD REFERENCE MATERIAL 1010a
(ANSI and ISO TEST CHART No. 2)

DOCUMENT RESUME

ED 241 566 TM 840 090

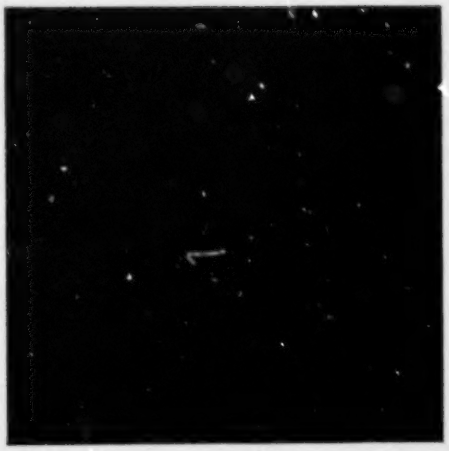
AUTHOR Levine, Michael V.; Drasgow, Fritz
TITLE Appropriateness Measurement with Aptitude Test Data and Estimated Parameters.
INSTITUTION Educational Testing Service, Princeton, N.J.; Graduate Record Examinations Board, Princeton, N.J.
REPORT NO GREB-75-3P
PUB DATE Mar 80
NOTE 43p.
AVAILABLE FROM Educational Testing Service, Publications Order Services, Dept. I-101, Princeton, N.J. 08541.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Aptitude Tests; Cheating; Error of Measurement; Error Patterns; Estimation (Mathematics); Higher Education; *Latent Trait Theory; Mathematical Models; *Multiple Choice Tests; *Testing Problems
IDENTIFIERS Outliers; *Test Appropriateness; Three Parameter Model

ABSTRACT

Appropriateness measurement is a general approach to the problem caused by multiple choice tests failing to measure accurately the ability of atypical examinees. The conceptual framework of appropriateness measurement is presented, and several statistical indices of the appropriateness of a multiple choice test for an examinee are noted. A series of interrelated experiments is used to demonstrate that appropriateness measurement is robust to errors in parameter estimation and to the presence of unidentified aberrant examinees in the test norming sample. Finally, studies with actual data are used to show that the three-parameter logistic model is sufficiently valid to support effective appropriateness measurement in applications with real examinees. (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****



GRE

GRADUATE RECORD EXAMINATIONS

APPROPRIATENESS MEASUREMENT WITH APTITUDE TEST DATA AND ESTIMATED PARAMETERS

Michael V. Levine

and

Fritz Drasgow
University of Illinois

GRE Board Professional Report GREB No. 75-3P

March 1980

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

EDUCATIONAL TESTING SERVICE, PRINCETON, NJ

ED241360

TM 840 090

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. Weidenmiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

APPROPRIATENESS MEASUREMENT WITH APTITUDE TEST DATA AND ESTIMATED PARAMETERS

Michael V. Levine

and

Fritz Drasgow
University of Illinois

GRE Board Professional Report GREB No. 75-3P

March 1980

Copyright © by Educational Testing Service. All rights reserved.

Abstract

Appropriateness measurement is a general approach to the problem caused by multiple choice tests failing to measure accurately the ability of atypical examinees. The conceptual framework of appropriateness measurement is presented, and several statistical indices of the appropriateness of a multiple choice test for an examinee are noted. A series of interrelated experiments is used to demonstrate that appropriateness measurement is robust to errors in parameter estimation and to the presence of unidentified aberrant examinees in the test norming sample. Finally, studies with actual data are used to show that the three-parameter logistic model is sufficiently valid to support effective appropriateness measurement in applications with real examinees.

APPROPRIATENESS MEASUREMENT WITH APTITUDE TEST DATA AND ESTIMATED PARAMETERS

In a large test administration a few examinees may be so unlike other examinees that their multiple choice aptitude test scores have limited value as ability measures. A particularly transparent example is provided by a hypothetical low ability copier who copies half of his answers from a much more able neighbor. Other test anomalies include:

a. improperly coached examinees who are shown the answers to some items before the exam begins

b. examinees with high ability but atypical schooling or low English fluency

c. exceptionally creative examinees who discover novel interpretations for some items

d. examinees who are very conservative in their use of partial information

e. examinees who make alignment errors over a block of items, answering the ninth item in the tenth place, the tenth item in the eleventh place, etc.

In each of these cases it can be argued that (1) the test score is not an appropriate measure of ability and (2) the item-by-item pattern of answers may be recognizably unusual. For example, the hypothetical low ability copier seems likely to have relatively many easy items wrong and

hard items right. A second example of an inappropriate test score and an unusual answer pattern occuring together is provided by the hypothetical examinee with an alignment error. He or she will most likely have a block of consecutive items incorrect and an unusual answer pattern within the block.

In sum, a multiple choice aptitude test may be a dubious measure of ability due to any one of many (although possibly rare) causes. In at least some cases, the item-by-item response pattern may contain evidence of this fact. In this paper, we consider the problem of using answer patterns to recognize inappropriate test scores.

Appropriateness Measurement: Its Objectives and Limitations

Appropriateness measurement is a general approach to the problem caused by inappropriate test scores. Its purpose is simply to identify inappropriate test scores. It is limited to cases, like those noted above, in which inappropriate test scores and unusual answer patterns tend to co-occur. Appropriateness measurement is implemented by statistics, called appropriateness indices, that measure the degree to which an examinee's answer pattern is "unusual", i.e., unlike the pattern expected from typical examinees.

In appropriateness measurement studies, examinees are sorted into two groups: (1) examinees with very unusual answer patterns as indicated by very extreme index values and (2) other examinees, i.e., examinees with typical index values. Appropriateness measurement is successful to the extent that the group of examinees with extreme index

values has a larger proportion of examinees with inappropriate scores than the group with typical index values.

Background

The first large scale appropriateness measurement study was reported by Levine and Rubin (Note 1). This study, reviewed below, provides the background and context for the theoretical developments and empirical results reported in this paper.[1]

Levine and Rubin (Note 1) identified three types of appropriateness indices and reported positive empirical findings with these indices. However, the generality of their findings is limited by properties of their data set. In particular, their data were simulated, the simulation parameters were available for use in defining appropriateness indices and aberrant examinees were unequivocally identified.

In this paper we use actual and simulated data to attack three problems raised by the Levine and Rubin study, namely:

1. Estimated vs. Known Item Parameters: With simulated data, item parameters (item difficulties, etc.) are known and need not be estimated prior to computing appropriateness indices. With actual data, parameters must be estimated. How seriously will appropriateness measurement be affected by error of estimation?

2. Unidentified Aberrants: In a simulation study, atypical examinees are unequivocally identified and a sample of truly normal examinees is available for item parameter estimation. With actual data, an unknown proportion of unidentified aberrants may be included in each large sample of nominally normal examinees. How will the presence of these aberrants affect parameter estimation and, consequently, appropriateness measurement?

3. Model Validity: Simulated data conform precisely to the psychometric model used to generate data and to formulate appropriateness indices. But there will be reliable contradictions to the assumptions of any tractable, nontrivial psychometric model in a large sample of actual data. Are currently available psychometric models sufficiently valid to support appropriateness measurement with actual data?

We will present results showing that our appropriateness indices (1) are not seriously degraded by the use of estimated parameters and (2) are not seriously degraded even when a relatively large proportion of aberrant examinees is initially (and improperly) treated as normal. Finally, we shall demonstrate that (3) detection rates with actual test data are comparable to detection rates with simulated data.

Review of Test Theories and Basic Appropriateness Measurement

Appropriateness measurement involves a two stage process: (1) a test norming or item parameter estimation stage followed by (2) a person measurement or index computation stage. This distinction parallels the seperation of parameters in latent trait models into (1) item or test characterizing parameters (such as item difficulties) and (2) individual difference or person characterizing parameters (such as abilities).

In the studies to be reported here, the test norming stage is developed around what we shall call the standard model. This test model is a version of the three parameter logistic model of item response theory (Birnbaum, 1968). According to our standard model, an answer sheet is generated by a two stage experiment. In the first stage an ability, θ, is sampled. In the second stage a sequence of realizations of independent, binary random variables, $u_1, u_2, \ldots, u_i, \ldots, u_n$ is generated. These are the item scores coded from the observed answer sheet with "1" denoting a correct response and "0" for an incorrect response. (The ability, θ, is not observed and the distribution of abilities is neither specified nor estimated in our current work.)

After some notation is introduced, the essential features of the standard model can be summarized with two equations. Let $\underset{\sim}{U}^{(j)}$ denote the jth examinee's answer pattern. Thus $\underset{\sim}{U}^{(j)}$, $j = 1, 2, \ldots, N$, is the vector of item scores,

$$\underset{\sim}{U}^{(j)} = \langle u_1^{(j)}, u_2^{(j)}, \cdots, u_n^{(j)} \rangle$$

for a test composed of n items. Let $P_i(\theta)$ denote the regression of the ith item score on ability, i.e., $P_i(\theta) = E(u_i|\theta)$. For a given θ, $P_i(\theta)$ can be interpreted as the conditional probability of an examinee, randomly selected from all examinees with ability θ, correctly answering the ith item. Then the assumption of a unidimensional latent space implies conditional independence of item scores. This is expressed by

$$(1) \quad \text{Prob}\{\underset{\sim}{U}^{(j)} | \theta^{(j)} = \theta\} = \prod_{i=1}^{n} P_i(\theta)^{u_i}[1 - P_i(\theta)]^{1-u_i} .$$

The independence of different examinees assumed in the model implies

$$(2) \quad \text{Prob}\{\underset{\sim}{U}^{(1)}, \underset{\sim}{U}^{(2)}, \ldots, \underset{\sim}{U}^{(N)} | \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(N)}\} = \prod_{j=1}^{N} \text{Prob}\{\underset{\sim}{U}^{(j)} | \theta^{(j)}\} .$$

To facilitate the test norming stage, each P_i is assumed to have the functional form

$$(3) \quad P_i(\theta) = c_i + (1 - c_i)\{1 + \exp[-Da_i(\theta - b_i)]\}^{-1}$$

for some positive a_i, real b_i and c_i with $0 \leq c_i \leq 1$. D is a conventional normalizing constant set equal to 1.702.

"Test norming" consists of estimating the numbers a_i, b_i and c_i. This is done by selecting a large sample of N presumably normal examinees, observing their answer patterns $U^{(1)}$, $U^{(2)}$, ..., $U^{(N)}$, and finding a set of a's, b's, c's and θ's that maximizes the likelihood

function in Equation 1. Our interest in the test norming stage is in the item parameters a, b and c; however, the θ's must be also be estimated in our current procedure.

Lord's LOGIST algorithm (Wood, Wingersky & Lord, Note 2; Wood & Lord, Note 3) can be used to maximize Equation 2. LOGIST is a widely used program, which can be easily obtained for replication of our experimental results. This program has been criticised (Wright, 1977); the relevance of our results to these criticisms is summarized in the last section of this paper.

The test norming stage can be thought of as specifying a continuum of models or statistical characterizations of typical examinees by using test responses of the first N nominally normal examinees in stage one to estimate a vector of item parameters. These parameter estimates can be substituted in Equation 2 to obtain an explicit formula for the likelihood of a new pattern of answers, say $U^{(N+1)}$. An intuition that has guided much of our research, and which will now be used to introduce the person measurement stage, is this: Suppose for all values of θ, $U^{(N+1)}$ appears improbable in the sense that $Prob(U^{(N+1)}|\theta)$ is very small. Then $U^{(N+1)}$ is badly fit by all models for individual data developed in the test norming stage.

In the person measurement stage the item parameters are treated as known and an index of goodness of fit is computed for each person's answer pattern. Our simplest index, L_0, is

$$L_0 = \log \max_{\theta} Prob(U^{(N+1)}|\theta) . \quad (4)$$

This index will be small if $\text{Prob}(U^{(N+1)}|\theta)$ is small for all values of θ. A small value of L_0 could result if many incorrectly answered easy items rule out high values of θ and many correctly answered hard items rule out low values of θ.

L_0 is our simplest appropriateness index. It detects aberrance surprisingly well. To improve upon it, we specify a model for aberrant data and two measures for the degree of aberrance in the next section.

Variable Ability Models and Appropriateness Indices

L_0, like χ^2, is sensitive to any type of badness of fit for the standard model. A generalization of the standard model is needed to detect aberrations of the specific kind that interest us.

In many of the most important types of test inappropriateness, the aberrant examinee behaves as if his ability were fluctuating from item to item. Thus, the low ability cheater appears to have a much higher ability for those items on which he has been coached. The high ability, low English fluency candidate behaves as if he had low ability on linguistically demanding items.

In the standard model, the examinee's ability, θ, is constant across items. In variable ability models, introduced by Levine and Rubin (Note 1, section 2), the examinee's ability is conceptualized as varying from item to item. For example, in the Gaussian model, each examinee is characterized by a pair of parameters: central ability θ_o and ability variance σ^2. Note that the standard model utilizes a single

individual difference parameter θ. According to the Gaussian model, the examinee has ability θ_1 on item 1, θ_2 on item 2, ..., and θ_n on item n. The model asserts that for given θ_0 and σ^2, the θ_i are independent, identically distributed normal or Gaussian random variables with mean = θ_0 and variance = σ^2. To specify the model more precisely we write and simplify the likelihood function $\text{Prob}(U|\theta_0,\sigma^2)$ that gives the conditional probability of a vector U of item responses:

$$(5)\quad \text{Prob}(\underset{\sim}{u}|\theta_0,\sigma^2) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\prod_{i=1}^{n}\left\{P_i(\theta_i)^{u_i}\left[1-P_i(\theta_i)\right]^{1-u_i}\phi(\theta_i;\theta_0,\sigma^2)d\theta_i\right\}$$

$$= \prod_{i=1}^{n}\int_{-\infty}^{\infty}P_i(t)^{u_i}\left[1-P_i(t)\right]^{1-u_i}\phi(t;\theta_0,\sigma^2)\,dt\ ,$$

where $\phi(t;\theta_0,\sigma^2)$ is the Gaussian density

$$\phi(t;\theta_0,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\frac{t-\theta_0}{\sigma}\right]^2\right\}\ .$$

Equation 5 is analogous to Equation 1 for the standard model.

To obtain a second appropriateness index testing for a specific departure from the standard model we compute the maximum likelihood ratio test statistic

$$(6)\qquad LR = L_n - L_0$$

where

$$L_n = \log \max_{\theta_0,\sigma} \text{Prob}(\underset{\sim}{u}|\theta_0,\sigma^2)$$

and

$$L_0 = \log \max_{\theta} \text{Prob}(\underset{\sim}{u}|\theta)\ ,$$

as before. LR measures the degree to which a variable ability model provides a better fit to the observed pattern of responses than the standard model.

The final appropriateness index or measure of goodness of fit is $\hat{\sigma}$, the maximum likelihood estimate of the ability standard deviation. This index is obtained by maximizing $\text{Prob}(U|\theta_o,\sigma^2)$ with respect to both θ_o and σ. The standard model is a special case of the Gaussian model in the sense that $\text{Prob}(U|\theta)$ is the limit of $\text{Prob}(U|\theta_o,\sigma^2)$ as σ decreases to zero. Consequently, a small $\hat{\sigma}$ can be interpreted as indicating a small degree of aberrance.

Index Evaluation and Roc Curves

Each application of appropriateness measurement will have different rewards for correctly identifying an inappropriate score and penalties for incorrectly classifying a normal examinee as aberrant. The ROC curve of statistical decision theory provides a graphic way to compare indices prior to the specification of rewards and penalties.

In applying an appropriateness index to classify examinees, a cut off or criterion value of the index is specified. For the present exposition, assume small index values indicate aberrance. At each criterion value, t, the proportion of aberrant examinees correctly identified as aberrant and the proportion of normal examinees improperly identified as aberrant can be denoted as

x(t) = proportion of normal examinees with index values $\leq t$

y(t) = proportion of aberrant examinees with index values $\leq t$.

An ROC curve results from plotting the $\langle x(t), y(t) \rangle$ pairs obtained for various values of t. A desirable ROC curve is one that rises sharply from the origin toward the upper left hand corner of the plot. In contrast, an ROC that lies along the diagonal of the plot indicates a random classification rule. That is, classifying examinees on the basis of flipping a coin yields a diagonal ROC. Clearly, an ROC curve which indicates an effective detection procedure is one which lies well above the diagonal.

Our results will be displayed via empirical ROC curves.

Simulation Procedures and Results

Levine and Rubin's (Note 1) simulation methods and results are reviewed here as we reanalyze some of their data and apply their methods to new data. To simulate a normal vector of item scores, they first sampled an ability θ from a normally distributed population with zero mean and unit variance. The examinee's first item score was generated by sampling a number uniformly distributed in the unit interval. If the sampled number was less or equal to $P_1(\theta)$ from Equation 3 then the first item was scored as correct; otherwise the item was scored as incorrect. The remaining item scores were obtained by independently drawing new uniformly distributed numbers and comparing them with $P_i(\theta)$ for i = 2, 3, ..., n.

The parameters, a_i, b_i, c_i, utilized in Equation 3 were those obtained by Lord's fitting of a three parameter logistic model to a large sample of Scholastic Aptitude Test, Verbal Section data (Lord, 1968). The actual simulation was implemented with Hambleton and Rovenelli's (1973) program. For technical details concerning the random number generators used, see Levine and Rubin (Note 1, appendix).

Aberrant examinees were simulated by modifying simulated normal answer sheets in various ways. In this paper we are primarily concerned with the "20% spuriously low" modification, but other modifications will also be reviewed briefly.

To create a spuriously high answer sheet, 20% of a normal simulated examinee's item responses are randomly sampled without replacement. If the sampled item was originally scored as correct, it is left unchanged. If the sampled item was not correct, it is rescored as correct.

To create a spuriously low examinee, 20% of a normal examinee's responses are sampled as before. But then a random number generator is used to simulate random guessing over the five multiple choice alternatives. No matter what the original sampled item score was, the sampled item is scored correct with probability 1/5 and scored as incorrect with probability 4/5.

Levine and Rubin observed that the modification of a normal answer pattern to create a spuriously low answer pattern frequently resulted in little or no change in the actual response vector. Clearly, if an aberrance producing process does not alter the objective response pattern, it is futile to attempt to detect the presence of the process

with an appropriateness index. Furthermore, if there is no effect on the objective response pattern and test score, there is little motivation for detecting aberrance. Consequently Levine and Rubin seperately analyzed the data from spuriously low examinees who had at least ten percent of their scores changed. Both correct responses changed to incorrect and incorrect responses changed to correct were counted toward the ten percent figure.

Levine and Rubin observed good aberrance detection with the total sample of spuriously low examinees and excellent detection for both the spuriously high examinees and the selected, large-score-change sample of spuriously low examnees. These generalizations held for all three indices.

There are two additional results obtained by Levine and Rubin that merit attention. First, Levine and Rubin systematically varied the percent of items modified, utilizing 4, 10, 20 and 40% treatments. As expected, they found increasing detectability as the percent of modified items was increased. A second finding by Levine and Rubin was that the three appropriateness indices, L_0, LR and $\hat{\sigma}$, yield quite similar patterns of detectability. Interestingly, no one index was substantially better or worse than the other two. Consequently, Levine and Rubin did not recommend usage of any one particular index in future research.

We shall reanalyze Levine and Rubin's data, using estimated item parameters instead of simulation parameters. We shall also apply appropriateness measurement techniques to actual test data.

To recapitulate and tabulate numerical details, we now summarize the Levine and Rubin data files relevant to this paper. They are:

1. NORMAL 3200: 3200 simulated answer sheets with normally distributed abilities; item parameters from Lord's (1968) fitting of the SAT-V; items scored either as correct or incorrect.
2. NORMAL 2800: The first 2800 records from NORMAL 3200.
3. LOW 200: Records 3001, 3002, ..., 3200 from NORMAL 3200 modified to simulate ability-unrelated responding on 20% of the test according to the 20% spuriously low modification.
4. LOW 102: One hundred and two records selected from LOW 200 that had at least 10% of the simulated examinee's original responses changed in the spuriously low modification.
5. HIGH 200: Records 2801, 2802, ..., 3000 from NORMAL 3200 modified to simulate cheating on 20% of the items according to the 20% spuriously high modification.

Study 1: Estimated Parameters

Problem: Levine and Rubin bypassed the first (test norming) stage of appropriateness measurement. Instead of estimating item parameters from a large sample of normal examinees, they used the true (simulation) parameters to compute appropriateness indices. Clearly, in an application with actual data, item parameters must be estimated. How sensitive are indices to item parameter estimation error? Can high detection rates be achieved with estimated parameters?

Method: Item parameters were estimated by applying Lord's maximum likelihood algorithm LOGIST to NORMAL 2800. L_0 was computed for each NORMAL 2800 examinee by evaluating the likelihood function at the LOGIST maximum likelihood estimate of ability. The LOGIST estimated item parameters were then used to compute L_0 for the LOW 102 response vectors by rerunning LOGIST for the LOW 102 data with all item parameters fixed at the values obtained from NORMAL 2800. In this way an estimated parameter L_0 appropriateness index value was obtained for each NORMAL 2800 and LOW 102 response vector.

Results: A close agreement between values of L_0 computed from exact parameter and estimated parameter index values was observed. In Figure 1 a bivariate scatterplot demonstrates this agreement for the critical LOW 102 sample. Each of the 102 simulated examinees contributes a point to the scatterplot. The x-coordinate of the point is L_0 computed with exact item parameters; the y-coordinate, with estimated item parameters. If there were perfect agreement between the two measures, the points would fall on the diagonal line, which has been drawn on the figure. We observe a very slight tendency for estimated index values to be smaller than exact index values for the most aberrant examinees in LOW 102 (L_0 less than -60).

Insert Figure 1 about here

The same close agreement was observed for the NORMAL 2800 sample. This is shown in Figure 2 for the first 100 simulated examinees in NORMAL 2800.

Insert Figure 2 about here

A more significant result is shown in the estimated parameter L_0 ROC curve presented in Figure 3. Note that detection rates are high, even for low false alarm rates. For example, 12.7 percent of the aberrant examinees can be correctly classified with an L_0 cut off score that does not misclassify a single NORMAL 2800 examinee. Further, 21.6 percent of the aberrant examinees are detected at a false alarm rate of .9 percent and 47.1 percent of the aberrants are detected at a 4.4 percent false alarm rate. Figure 3 is in close agreement with Levine and Rubin's exact parameter LR ROC curve computed from the same data (their Figure 8). In fact, there is a very small superiority for the estimated parameter ROC; using estimated parameters there are 0, .9 and 4.4 percent false alarms at hit rates of 11.8, 21.6 and 47.1 percent in comparison to .1, .9 and 4.8 percent false alarms at the corresponding points in the exact parameter ROC.

Insert Figure 3 about here

Discussion: Our results may initially seem surprising in view of Lord's (Note 4) study of the disparity between LOGIST estimated item parameters and simulation parameters.[2] Lord found much more variation around the diagonal in his bivariate plots of simulation item parameters vs. estimated item parameters than appears in our plots of estimated parameter L_0 vs. simulation parameter L_0. The discrepancy becomes less surprising when it is recalled that L_0 is the value of a likelihood function and Lord's parameter estimates are obtained by maximizing the likelihood function. If the likelihood function is relatively flat, then the maximizing values of the arguments of the function will be hard to determine precisely because a small parameter change results in a very small likelihood change. Somewhat paradoxically, flatness of the likelihood function can simplify the problem of calculating L_0, the value of the function near its maximum. The value of the likelihood function will be almost constant for parameter values in the neighborhood of the maximizing values.

A possible artifact complicating the interpretation of these results is seen by considering an analogy from multiple regression. The ROC curve will be high to the degree that the estimated parameters fit NORMAL 2800 better than LOW 102. NORMAL 2800 can be considered analogous to a multiple regression derivation sample and LOW 102 to a cross-validation sample. At least with small samples, overfitting is expected in the derivation sample and shrinkage in the cross-validation sample. It may be for this reason that L_0 - as a measure of goodness of fit - tends to be smaller in LOW 102. The fact that estimated parameter

aberrance scores for the most aberrant LOW 102 examinees are lower than exact parameter aberrance scores supports the suspicion of overfitting. But any overfitting should result in relatively high NORMAL 2800 scores (i.e., many points above the diagonal in Figure 2) and this was not found. Furthermore, the discrepancy between estimated and exact parameter index values is so small that abcissa values on the ROC curve would be changed less than .0004 if the 13 lowest LOW 102 index values fell exactly on the diagonal in Figure 1.

If overfitting were a significant artifact, then one would expect poor detection if either the normal group used to evaluate an index was distinct from the norming group utilized for item parameter estimation or if the aberrant and normal groups were pooled to form the norming sample. In Study 4 and in our current work (Drasgow, 1978), the norming and normal groups are distinct. In the next study to be reported here, the aberrant and normal groups are combined to form a single group used to estimate item parameters.

Study 2: Heterogeneous Norming Sample, Classification and Norming Sample Equal

Problem: In a simulation study all aberrant answer sheets are clearly identified since they are generated by the experimenter. In an actual study some undetected aberrants are likely to be included in the norming sample. Will unsuspected aberrants in the norming sample seriously degrade item parameter estimates and determine the person measurement stage of appropriateness measurement?

A secondary problem is related to the possible overfitting and shrinkage noted in Study 1. Individual simulation examinees are expendable. They can be used for test norming and ignored thereafter because new, statistically equivalent answer sheets can easily be generated for use in the person measurement stage. However, in actual studies sample sizes are fixed, and it generally will be important to evaluate appropriateness for the examinees in the norming sample. Will current estimation procedures "overfit" aberrant examinees in the norming sample, or will it be possible to use an appropriateness index to identify norming sample aberrants?

Method: NORMAL 2800 and LOW 200 were merged to form a data file with a large proportion of aberrant examinees. Item parameters were estimated using all 3000 simulated examinees. As before, L_0 was computed for all examinees by evaluating the likelihood function at the LOGIST maximum likelihood estimate of ability. New index values for LOW 102 were compared with exact parameter index values and the L_0 ROC curve was computed to evaluate detectability.

Results and Discussion: Figure 4 shows that estimating item parameters in a large sample with a large proportion of spuriously low examinees need not noticeably degrade appropriateness measurement. The simulation parameter L_0 ROC curve has hit rates of 11.7, 21.6 and 47.1 percent at false alarm rates of .1, .9 and 4.8 percent. The corresponding heterogeneous norming sample false alarm rates are .1, 1.2 and 4.9 percent. Clearly, the net effect on appropriateness measurement of

estimating item parameters from this heterogeneous sample is negligible.

Insert Figure 4 about here

Figure 5 contains the bivariate scatterplot of exact parameter L_0 values plotted against L_0 values computed from item parameters estimated in the heterogeneous sample. The relatively high frequency of points above the diagonal in Figure 5 indicates an overfitting effect for aberrant examinees. But both Figures 4 and 5 support the conclusion that the effect is small. Figure 5 does so because all points are tightly clustered about the diagonal, i.e., there is little difference between true parameter L_0 values and heterogeneous sample estimated parameter L_0 values. The high detectability exhibited in the ROC supports the contention that overfitting is small because a large overfitting effect would have reduced normal-aberrant group differences and therefore reduced detectability of aberrance.

Insert Figure 5 about here

We advise interpreting these results on overfitting cautiously. Parameters were estimated in a very large sample (N=3000). Further, the nature of spuriously low aberrance may be essential to the small effect. The distinction between bias and sampling error is useful in

understanding this point. The process we hypothesize to underlie spuriously lowness is essentially unsystematic; that is, atypical schooling, alignment errors and exceptional creativity lead to incorrect responses on different examination items. Thus the different examinees will tend to have competing effects on item parameter estimates. Consequently, the presence of aberrance in the norming group should affect the sampling error of item parameter estimates to some extent, but should have a relatively small effect on the bias of the estimates. In addition, the effect on the sampling error is tolerable due to the large sample size.

Study 3: Actual SAT-V Data

Problem: The one dimensional, three parameter logistic model is the most general model for which there is a well-developed parameter estimation literature. It is easy to formulate refinements and generalizations of this model, but this would require a long and costly research program to develop parameter estimates. Is the logistic model sufficiently descriptive to detect spuriously low examinees in actual test data?

Method: Three thousand "low omitting rate" examinees were sampled from an administration of the ETS Scholastic Aptitude Test-Verbal Section (SAT-V). All 3000 examinees responded to at least 90 percent of the test items.

LOGIST was used to estimate item parameters from these 3000 nominally normal examinees. A file of 200 aberrant examinees was then created by applying the 20 percent spuriously low modification to answer sheets from examinees 2801, 2802, ..., 3000.

L_0 was computed by maximizing the individual likelihood functions for the three parameter logistic model by Newton-Raphson iterations. Then the algorithm written by Gruvaeus and Jöreskog (Note 5) was used to maximize the likelihood of each examinee's item responses for the Gaussian model. Finally, the likelihood ratio statistic in Equation 6 was computed for each examinee by comparing the log likelihood of an answer sheet in the Gaussian model (L_n) to the log likelihood for the standard model (L_0). Item parameters estimated in the sample of 3000 nominally normal examinees were used to compute L_0 and L_n. In computing L_0 and L_n, omitted and not reached items were ignored.

Results: Figures 6 and 7 present the ROC curves for the L_0 and LR appropriateness indices, respectively. It is apparent that high detection rates are obtained at low false alarm rates. In particular, there are .2, .8 and 3.3 percent false alarms at hit rates of 10, 20 and 40 percent for the L_0 index and the LR index yields .04, .8 and 4.3 percent false alarms at the same hit rates. These results are even more impressive when it is noted that the aberrant group consists of all 200 spuriously low examinees; no spuriously low examinee was deleted from the analysis due to an insufficient change in item scoring. In fact, 42 examinees had seven or fewer item responses changed (from correct to

incorrect and incorrect to correct) when subjected to the spuriously low modification.

Insert Figures 6 and 7 about here

Discussion: It might be argued that the ROC curves in Figures 6 and 7 capitalize on statistically overfitting data from the normal examinees. This argument rests on the fact that the 2800 examinees constituting the normal sample were included in the norming sample whereas item responses from the aberrant examinees were included in the norming sample prior to the spuriously low modification. That is, the post-tampering response vectors were not included in the norming sample for the spuriously low group. However, the overfitting problem appears inconsequential for two reasons. First, as seen in Studies 1 and 2, overfitting does not seem to pose serious difficulties for appropriateness measurement when spurious lowness is unsystematic and item parameters are estimated in a large sample. Second, note that the majority of the item responses made by spuriously low examinees did contribute to item parameter estimation: No more than 20 of the 85 items were modified. Consequently, we believe that overfitting had very little, if any, effect on the detection of aberrance in Figures 6 and 7.

It is interesting to compare the effectiveness of the L_0 and LR indices in detecting aberrance. Clearly, in Figures 6 and 7 the L_0 index is superior to the LR index at high false alarm rates. But we expect that there are many more normal examinees than aberrant examinees. Hence the base rate difference causes the discrepancy between the L_0 and LR ROC curves at moderate to high false alarm rates to be unimportant; index performance is crucial only at very low false alarm rates. The LR index detects 19 aberrant examinees (out of 200) without misclassifying a single normal examinee whereas L_0 detects only 6 without a misclassification. At a false alarm rate of .5 percent, LR detects 37 aberrant examinees and L_0 detects 32. Thus at very low false alarm rates, there is some evidence in favor of the LR index, but this evidence is not especially compelling.

Study 4: Actual GRE-V Data

Problem: Levine and Rubin (Note 1) investigated the effectiveness of appropriateness measurement using data that simulated the SAT-V exam. Studies 1 and 2 presented here further analyzed the data generated by Levine and Rubin. In Study 3, an actual SAT-V data set was examined. A common theme across all these studies is that actual or simulated SAT-V data sets were investigated. In Study 4, we examine actual Graduate Record Examination - Verbal (GRE-V) data. Will the effectiveness of appropriateness measurement be diminished when the methods, developed using simulated and actual SAT-V data, are applied to GRE-V data?

Method: The responses to the GRE-V by 10,000 examinees were utilized in the following manner. First, a file of 3000 examinees (FILE1) with a wide range of ability and unrestricted omitting was created by selecting examinees 1,2,3,11,12,13, ..., 9991,9992,9993. This data set was then analyzed by LOGIST to obtain item parameter estimates. A second file was created by examining the item responses of the remaining 7000 examinees and selecting examinees with a low omitting rate. A total of 2470 examinees who had answered at least 86 of the 95 GRE-V items was obtained. Two hundred of these examinees were subjected to the 20 percent spuriously low modification. These modified response vectors formed the aberrant group for Study 4 and the remaining 2270 response vectors formed the normal group.

L_0 was computed for the 200 aberrant and 2270 normal response vectors as in Study 3 using the FILE1 item parameter estimates. Notice that here the test norming sample is distinct from the normal and aberrant samples.

Results: Figure 8 presents the ROC curve for the L_0 index. Clearly, detection rates are quite high. At hit rates of 10, 20 and 40 percent, there are .3, 1.4 and 3.7 percent false alarms. These detection rates are not substantially different from those obtained in Study 3.

Insert Figure 8 about here

Discussion: The results of Study 4 are important for two reasons. First, the criticism of overfitting the normal sample, which could be made for Studies 1 and 3, is not relevant for Study 4. Because the test norming sample (FILE1) was composed of examinees not included in the normal and aberrant groups, no differential statistical overfitting of the data for the normal and aberrant groups is possible. Consequently, the results of Study 4 can be interpreted unambiguously.

The finding that appropriateness measurement is effective for GRE-V data is important for a second reason. Levine and Rubin (Note 1) originally considered a variety of appropriateness indices. After examining the effectiveness of each index using simulated SAT-V data, Levine and Rubin selected the most effective indices for further study. The extent to which the most effective indices were capitalizing on test characteristics unique to the SAT-V was unknown. Study 4 shows that the methods developed using SAT-V data are sufficiently general that they can be implemented for GRE-V data.

Reference Notes

1. Levine, M. V. & Rubin, D. B. Measuring the appropriateness of multiple-choice test scores. Research Bulletin 75-31. Princeton, N.J.: Educational Testing Service, 1975.

2. Wood, R. L., Wingersky, M. S. & Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. Research Bulletin 76-6. Princeton, N.J.: Educational Testing Service, 1976.

3. Wood, R. L. & Lord, F. M. A user's guide to LOGIST. Research Bulletin 75-4. Princeton, N.J.: Educational Testing Service, 1975.

4. Lord, F. M. Evaluation with artifical data of a procedure for estimating ability and item characteristic curve parameters. Research Bulletin 75-33. Princeton, N.J.: Educational Testing Service, 1975.

5. Gruvaeus, G. T. & Jöreskog, K. G. A computer program for minimizing a function of several variables. Research Bulletin 70-14. Princeton, N.J.: Educational Testing Service, 1970.

References

Donlon, T. F. & Fischer, F. E. An index of an individual's agreement with group determined item difficulties. Educational and Psychological Measurement, 1968, 28, 105-113.

Drasgow, F. Statistical indices of the appropriateness of aptitude test scores (Doctoral dissertation, University of Illinois, 1978). Dissertation Abstracts International, 19xx, xx, xxxxx. (University Microfilms No. xx-xxxx)

Flier, H. van der. Environmental factors and deviant response patterns. In Y. H. Portinga (Ed.), Basic problems in cross-cultural psychology. Amsterdam: Swets and Zeitlinger, 1977.

Ghiselli, E. E. Differentiation of individuals in terms of their predictability. Journal of Applied Psychology, 1956, 40, 374-377.

Ghiselli, E. E. Differentiation of tests in terms of the accuracy with which they predict for a given individual. Educational and Psychological Measurement, 1960, 20, 675-684. (a)

Ghiselli, E. E. The prediction of predictability. Educational and Psychological Measurement, 1960, 20, 3-8. (b)

Hambleton, R. K. & Rovenelli, R. A. A FORTRAN IV program for generating response data from logistic test models. Behavioral Sciences, 1973, 18, 74.

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Footnotes

This research was primarily supported by funds provided by the Educational Testing Service and grants from the Graduate Record Examination Board (No. 75-3) and the College Entrance Examination Board (PJ 82201) to Michael V. Levine. The points of view and opinions expressed here do not represent the position or policy of ETS, GREB or CEEB.

The research reported here was initiated by the senior author while he was at ETS. Much of this article was written while the junior author was at the University of Illinois.

[1]For independent, contemporary work that appears similar in conception, see van der Flier (1977). For possibly related research based on classical test theory, see Ghiselli (1956; 1960a; 1960b) and Donlon and Fischer (1968).

[2]LOGIST version 2.B was used in the present research. LOGIST has since been modified.

Figure Captions

Figure 1. Bivariate scatterplot of L_0 computed from exact and estimated (from NORMAL 2800 data) item parameters for LOW 102 response vectors.

Figure 2. Bivariate scatterplot of L_0 computed from exact and estimated (from NORMAL 2800 data) item parameters for the first 100 simulated examinees in NORMAL 2800.

Figure 3. ROC curve describing detection of LOW 102 response vectors by L_0 index computed from item parameters estimated in NORMAL 2800.

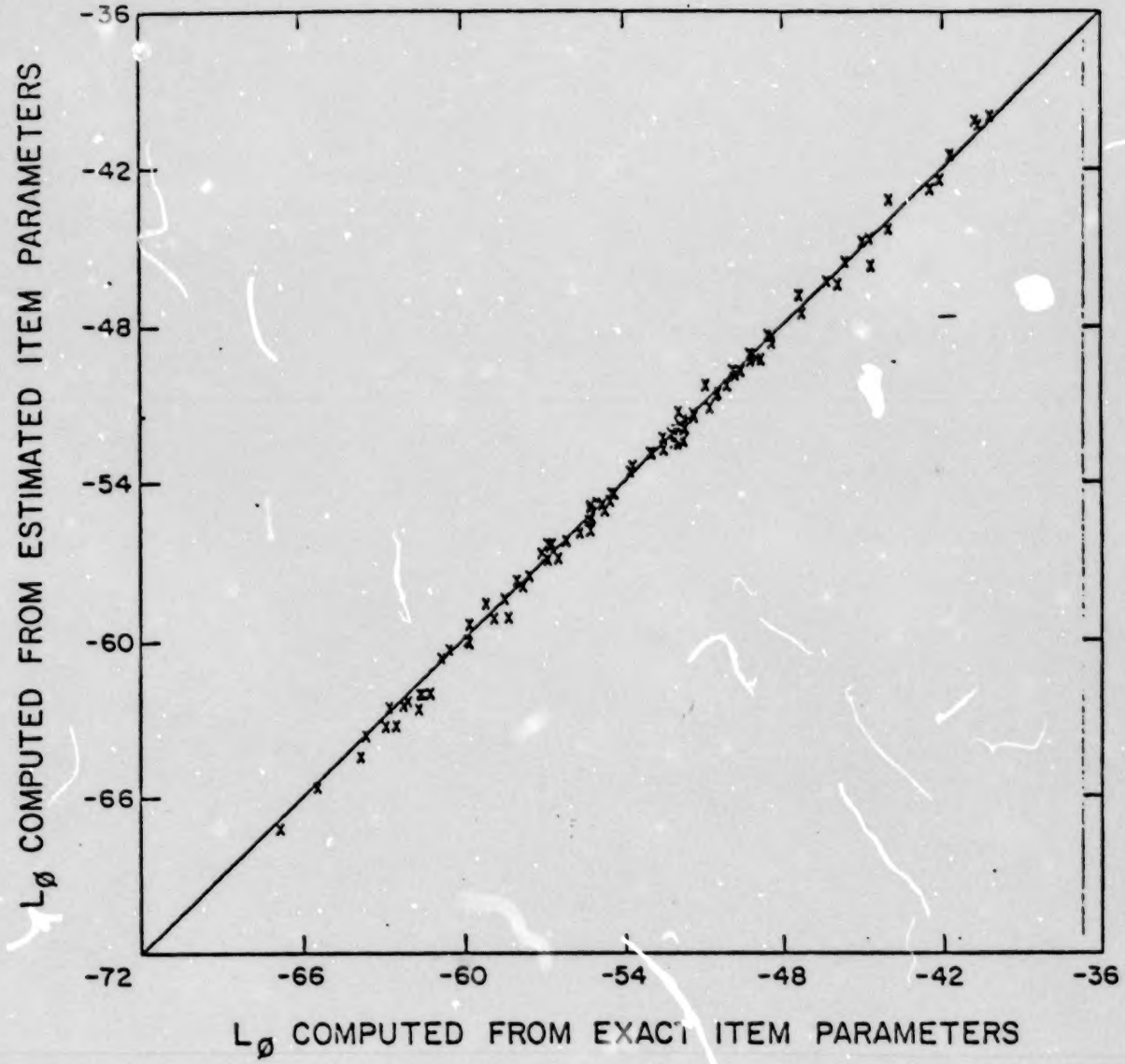
Figure 4. ROC curve describing detection of LOW 102 response vectors by the L_0 index computed from item parameters estimated in NORMAL 2800 and LOW 200 data.

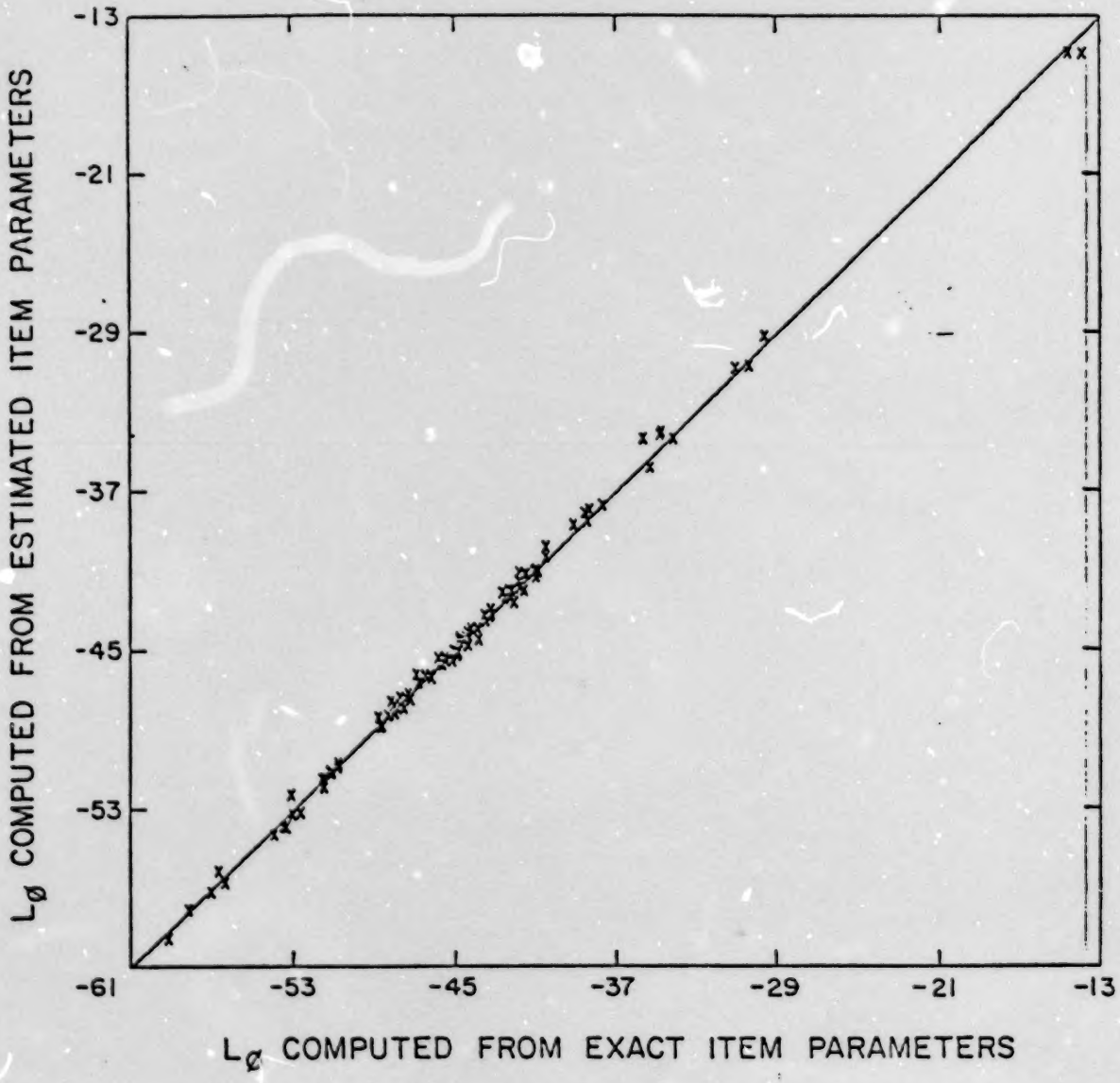
Figure 5. Bivariate scatterplot of L_0 computed from exact and estimated (from NORMAL 2800 and LOW 200 data) item parameters for LOW 102 response vectors.

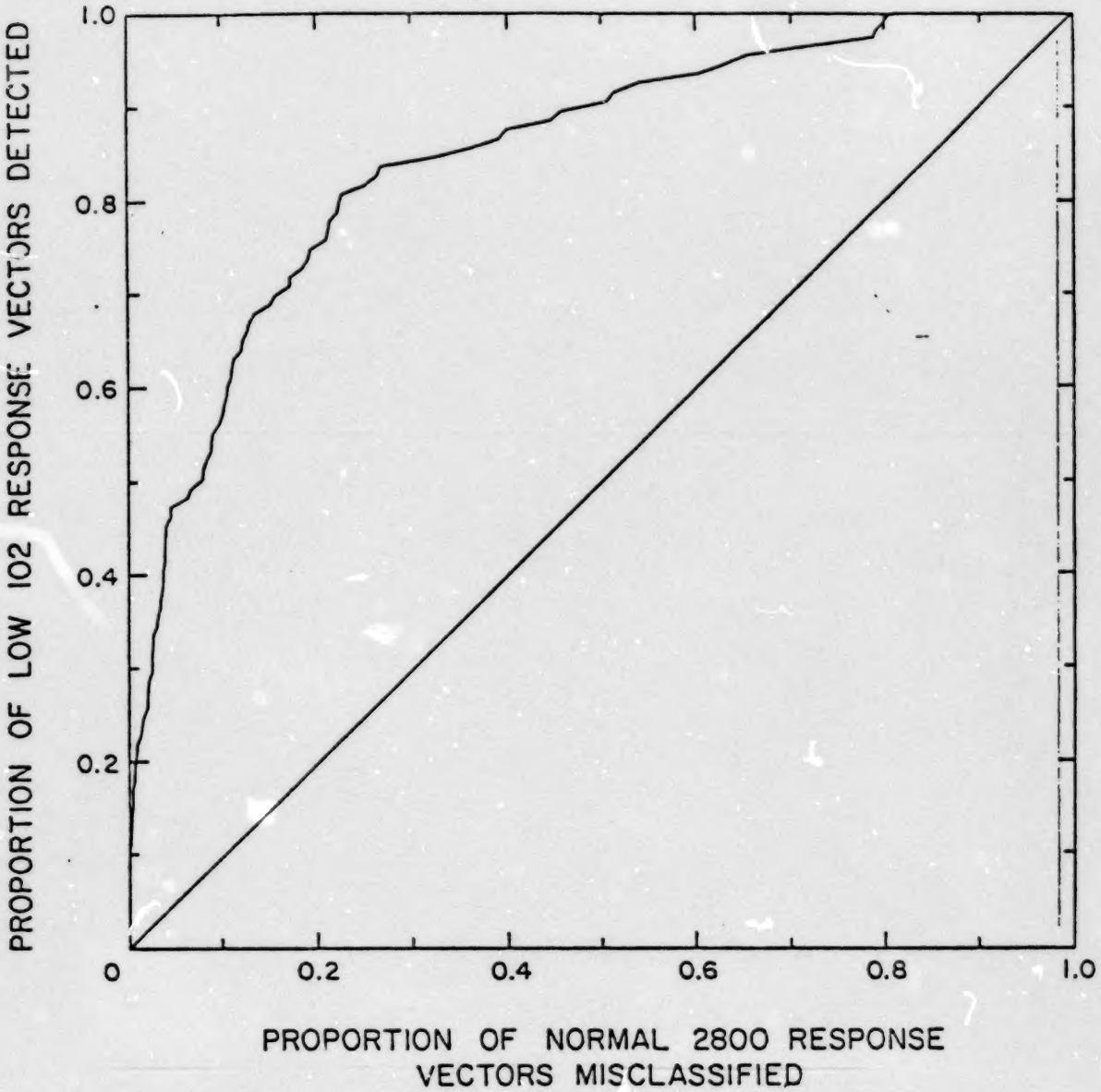
Figure 6. ROC curve describing detection of spuriously low response vectors by L_0 index for actual SAT-V data.

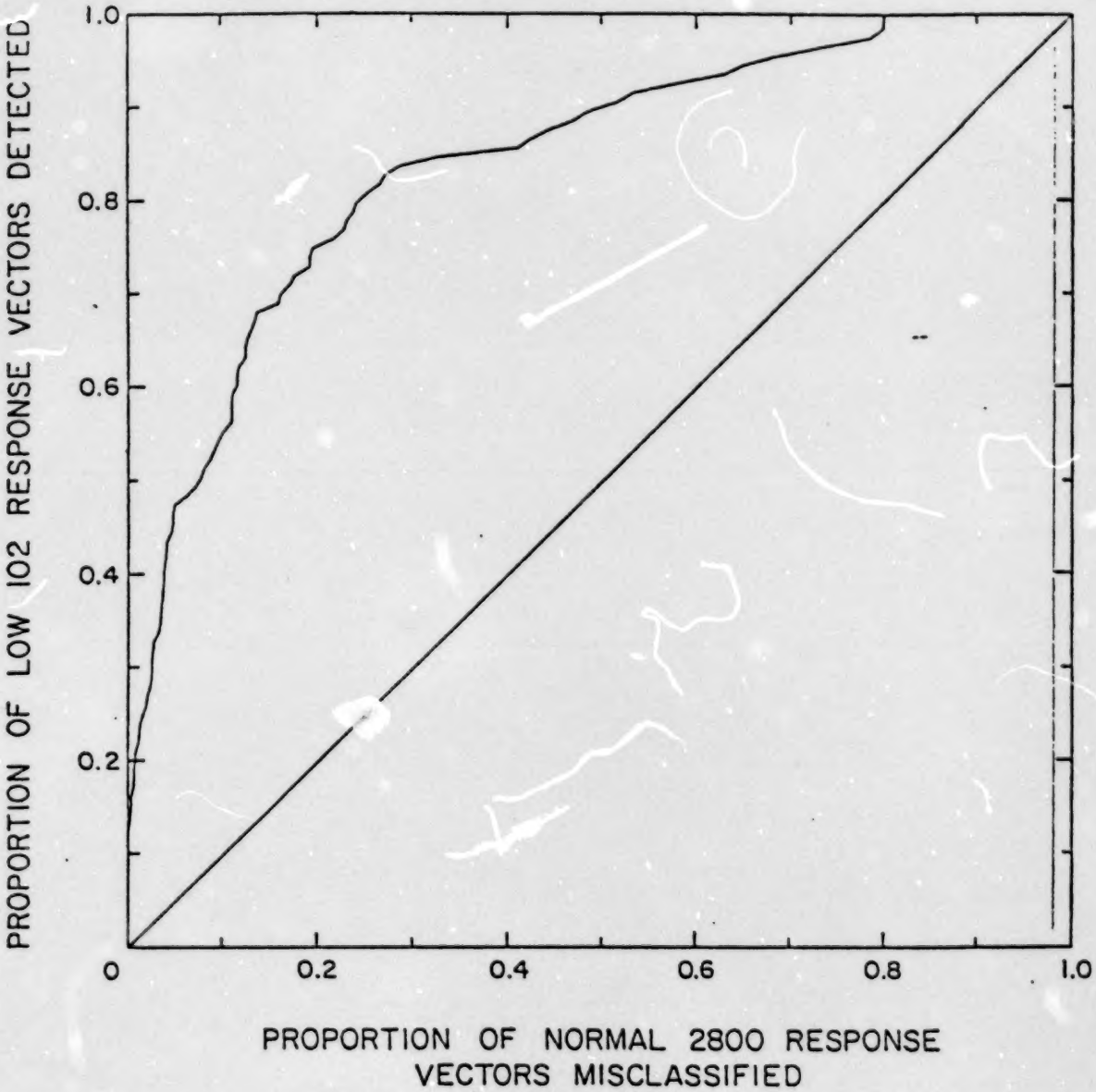
Figure 7. ROC curve describing detection of spuriously low response vectors by LR index for actual SAT-V data.

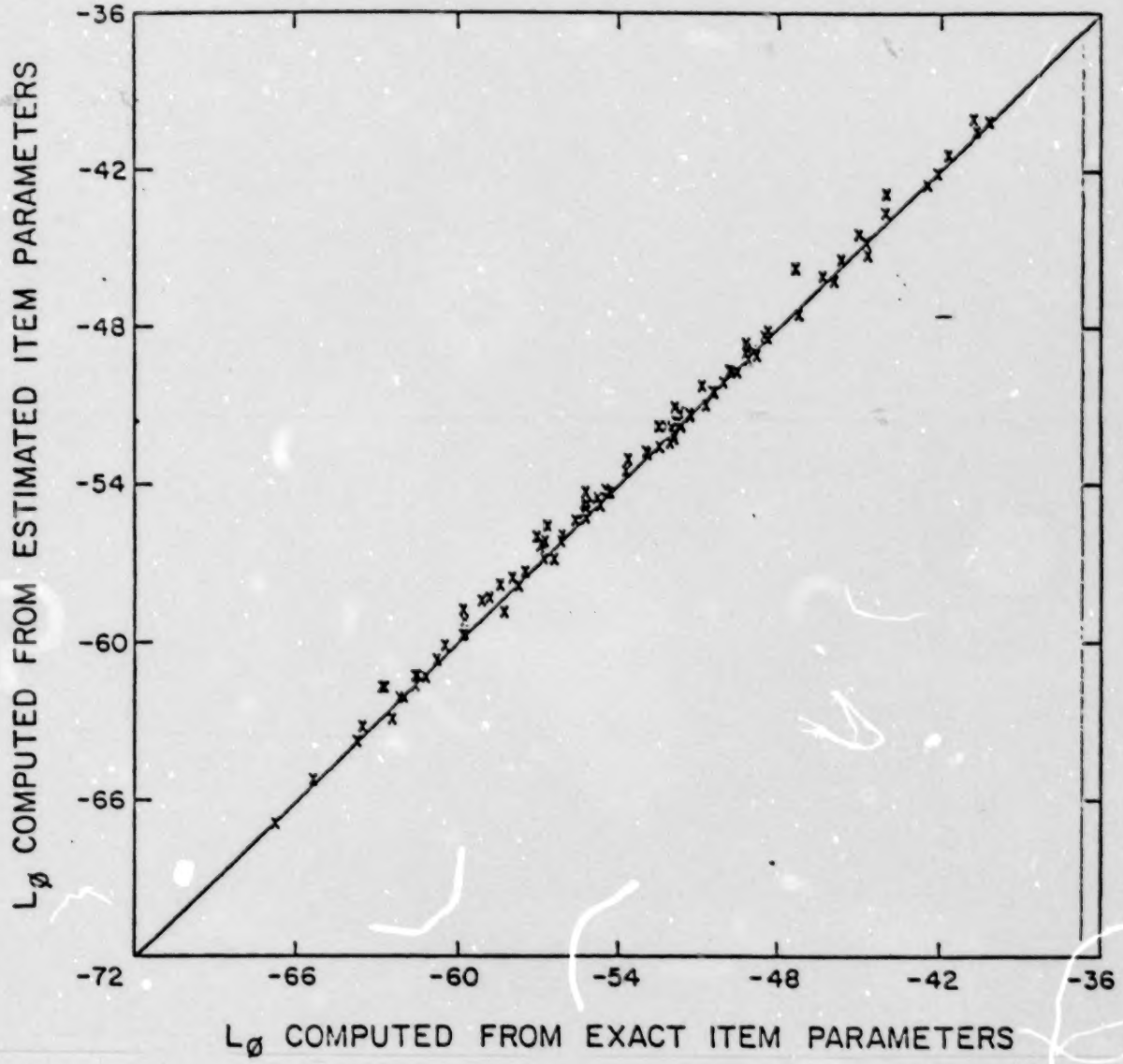
Figure 8. ROC curve describing detection of spuriously low response vectors by L_0 index for actual GRE-V data.

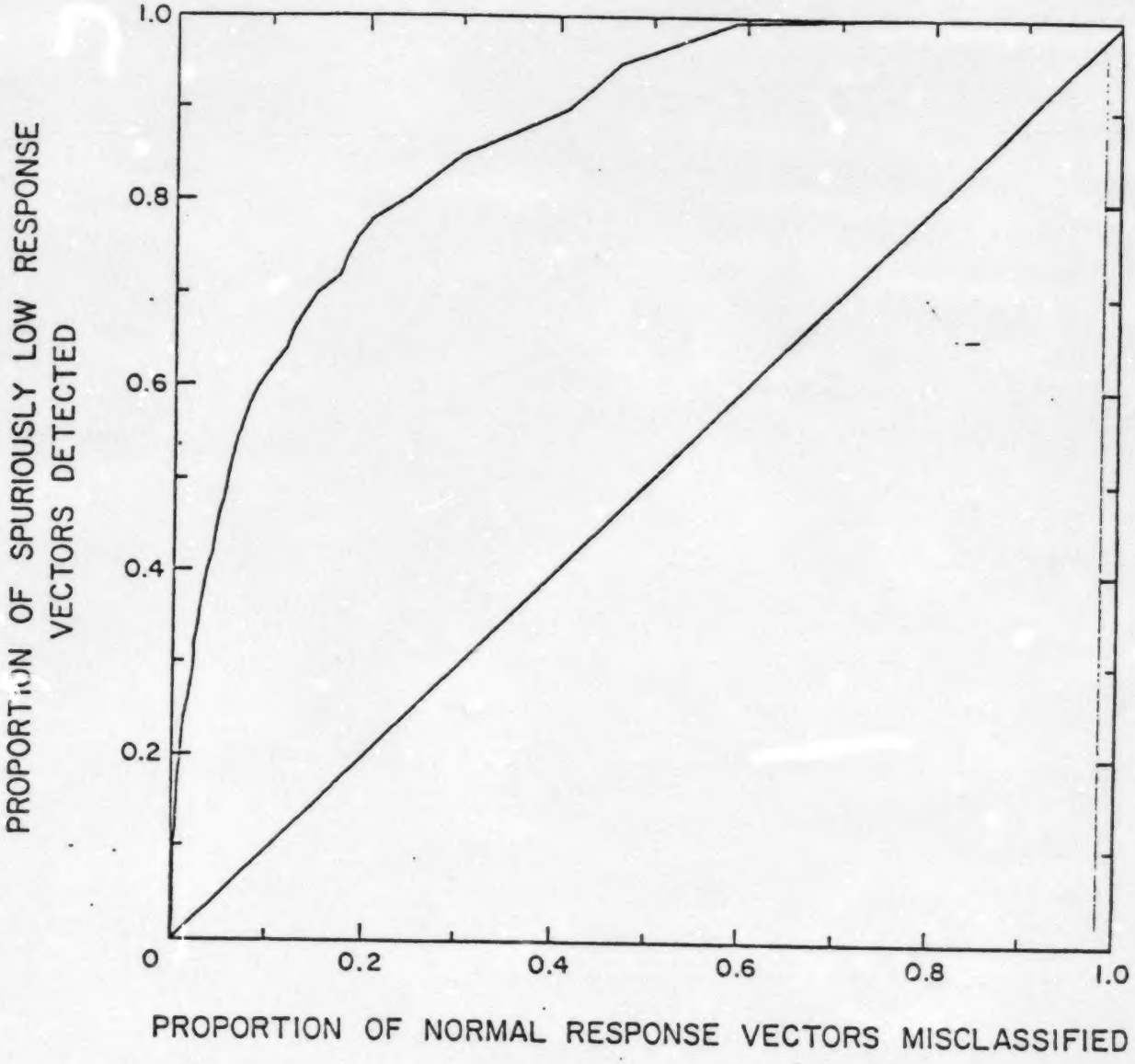

-36
-42
-48
-54
-60
-66
-72
-66
-60
-54
-48
-42
-36
L_ϕ COMPUTED FROM ESTIMATED ITEM PARAMETERS
L_ϕ COMPUTED FROM EXACT ITEM PARAMETERS

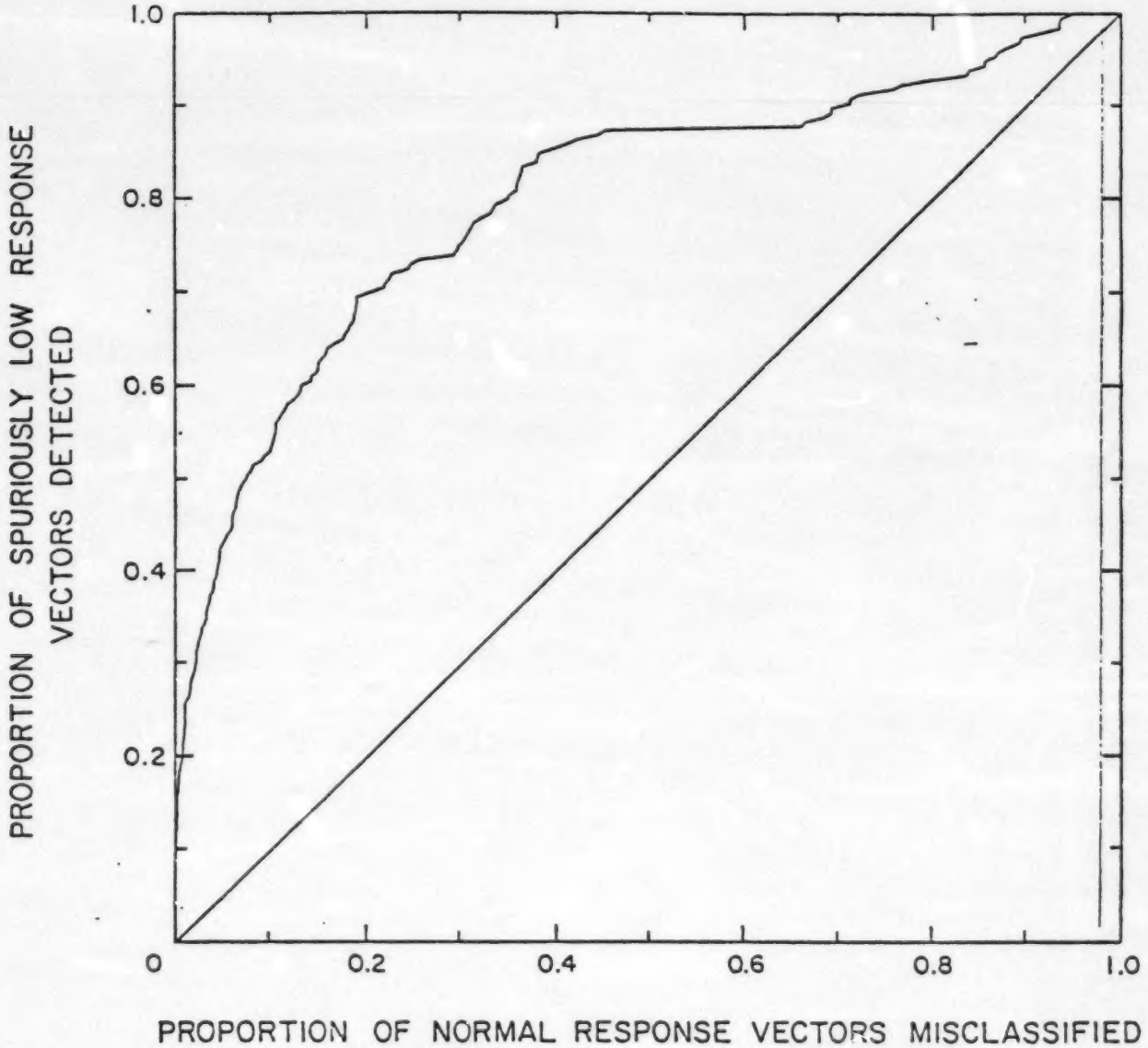

-13
-21
-29
-37
-45
-53
-61
-53
-45
-37
-29
-21
-13
Lø COMPUTED FROM ESTIMATED ITEM PARAMETERS
Lø COMPUTED FROM EXACT ITEM PARAMETERS

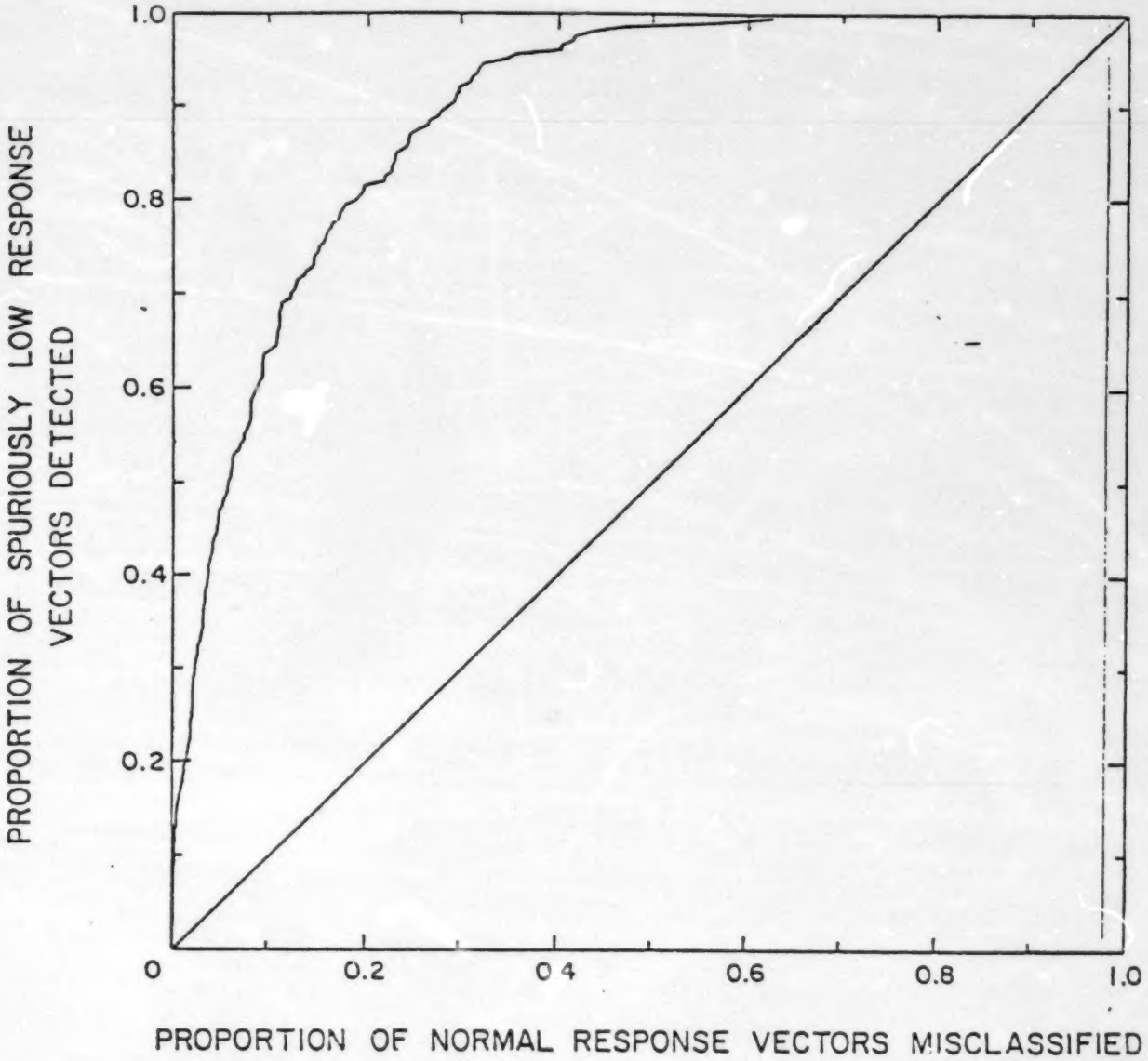

PROPORTION OF LOW 102 RESPONSE VECTORS DETECTED
1.0
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1.0
PROPORTION OF NORMAL 2800 RESPONSE VECTORS MISCLASSIFIED


PROPORTION OF LOW 102 RESPONSE VECTORS DETECTED
1.0
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1.0
PROPORTION OF NORMAL 2800 RESPONSE VECTORS MISCLASSIFIED


-36
-42
-48
-54
-60
-66
-72
-66
-60
-54
-48
-42
-36
Lø COMPUTED FROM ESTIMATED ITEM PARAMETERS
Lø COMPUTED FROM EXACT ITEM PARAMETERS


1.0
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1.0
PROPORTION OF SPURIOUSLY LOW RESPONSE VECTORS DETECTED
PROPORTION OF NORMAL RESPONSE VECTORS MISCLASSIFIED


PROPORTION OF SPURIOUSLY LOW RESPONSE VECTORS DETECTED
1.0
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1.0
PROPORTION OF NORMAL RESPONSE VECTORS MISCLASSIFIED


PROPORTION OF SPURIOUSLY LOW RESPONSE VECTORS DETECTED
1.0
0.8
0.6
0.4
0.2
0
0.2
0.4
0.6
0.8
1.0
PROPORTION OF NORMAL RESPONSE VECTORS MISCLASSIFIED